Nucleic Acids Research, 2004, Vol. 32, Database issue ***D323–D325***
DOI: 10.1093/nar/gkh135

yMGV: a cross-species expression data mining tool

Gaëlle Lelandais[1,3], Stéphane Le Crom[2], Frédéric Devaux[1], Stéphane Vialette[1], George M. Church[4], Claude Jacq[1] and Philippe Marc[4,*]

[1]Laboratoire de Génétique Moléculaire, CNRS UMR8541 and [2]Laboratoire de Biologie Moléculaire du Développement, INSERM U368, Ecole Normale Supérieure, 46 Rue d'Ulm, 75005 Paris, France, [3]Equipe de Bioinformatique Génomique et Moléculaire, INSERM E346 Université Paris 7, case 7113, 2 place Jussieu, 75005 Paris, France and [4]Lipper Center for Computational Genetics and Department of Genetics, Harvard Medical School, 77 Louis Pasteur Avenue, Boston, MA 02115, USA

Received September 14, 2003; Revised and Accepted October 27, 2003

ABSTRACT

The yeast Microarray Global Viewer (yMGV @ http://transcriptome.ens.fr/ymgv) was created 3 years ago as a database that houses a collection of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* microarray data sets published in 82 different articles. yMGV couples data mining tools with a user-friendly web interface so that, with a few mouse clicks, one can identify the conditions that affect the expression of a gene or list of genes regulated in a set of experiments. One of the major new features we present here is a set of tools that allows for inter-organism comparisons. This should enable the fission yeast community to take advantage of the large amount of available information on budding yeast transcriptome. New tools and ongoing developments are also presented here.

INTRODUCTION

Although several databases have been created to manage published microarray data, many of the associated tools are underutilized due to cumbersome user interfaces and non-intuitive output. This is the most common difficulty confronting the mining and visualization of the vast amount of data produced by genomic technologies. The yeast Microarray Global Viewer (yMGV) is a data mining tool coupled to a multi-organism database that currently houses *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* expression data. The philosophy of yMGV is to empower biologists by providing a straightforward data mining interface, and by generating easily interpretable, mostly graphical, output. This tool has matured since its creation in 2001, and is now recognized as an exemplary approach to the retrieval and interpretation of valuable biological information (1–3).

The basic features of yMGV have been described previously (4); here we present recent improvements to the data set and interface, and plans for the development of future modules.

EVOLUTION OF THE DATA SET

New data have been added to the database on a regular basis since the release of the original version, which contained 39 microarray data sets. Today, the yMGV database contains data from 1544 genome-wide expression experiments, representing 82 microarray publications. Importantly, expression data sets from the *S.pombe* Sanger Institute project (5) have been included in version 2, enabling inter-organism queries. The database architecture is designed to allow the addition of data pertaining to other organisms in the near future.

EVOLUTION OF TOOLS

yMGV is under continuous development, and version 2 has been available since April 2003. Recent improvements allow the user to critically assess the published data, e.g. summary statistics, such as the mean and standard deviation of the $\log_2$(ratio) distribution of a given microarray data set, are reported along with the $\log_2$(ratios) in that data set. New links to external databases have been added, and their connections improved (see Supplementary Material for current URLs). It is now possible to directly post a list of genes generated using yMGV to other online tools, e.g. to KEGG for metabolic mapping (6), to RSA tools for *cis*-regulatory motif discovery (7) or to SGD for Gene Ontology (GO) term mapping (8).

Several additional features are entirely new to the database. Two of them, cross-species transcriptome comparison and compendium modules, are detailed below.

CROSS-SPECIES TRANSCRIPTOME COMPARISON

Since its inception, a major goal of yMGV has been to incorporate data originating from different organisms (9), and the database schema has been designed to accommodate any genome described using GO formalisms (8) (see logical scheme in Supplementary Material). Toward this goal, the second organism to have been added to yMGV is the fission yeast *S.pombe*.

Intra-species data analyses carried out by yMGV have been extended to incorporate inter-species data, allowing comparisons of gene expression between orthologs. To facilitate

*To whom correspondence should be addressed. Tel: +1 617 432 4136; Fax: +1 617 432 7266; Email: pmarc@genetics.med.harvard.edu

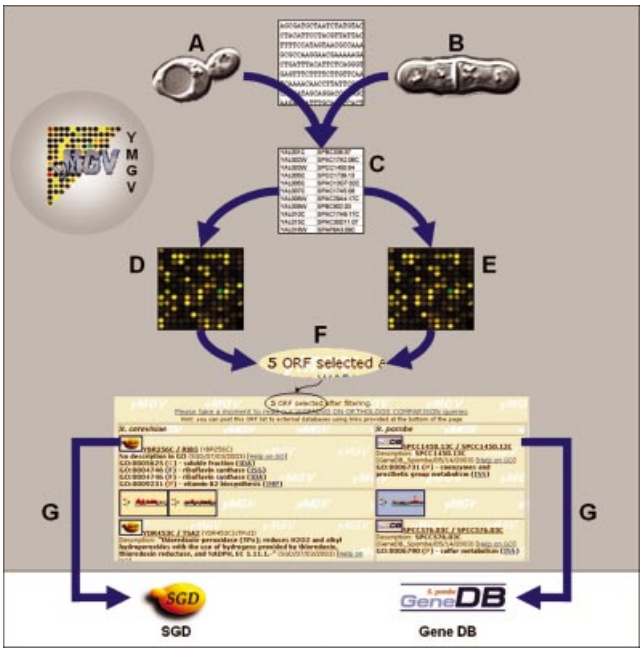

Figure 1. Using yMGV to compare gene expression in two organisms. yMGV allows comparison of gene expression of two organisms (**A** and **B**) (currently *S.cerevisiae* and *S.pombe* only) using an orthology table (**C**) constructed using sequence information. The user can apply filters to the transcriptome of one or both organisms (**D** and **E**) and get the list of orthologous gene pairs that fit the required expression profile and satisfy the filter parameters (**F**). GO description and links to organism-specific databases are provided for each gene (**G**).

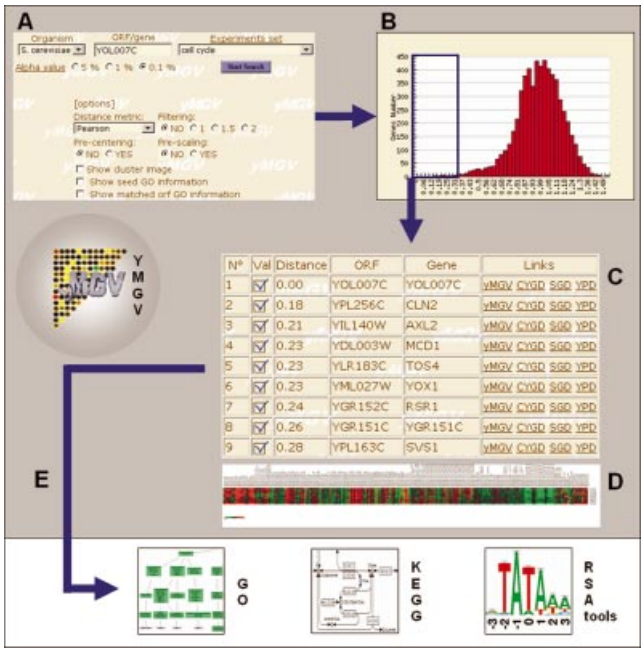

Figure 2. Using yMGV to find genes co-expressed in a subset of conditions. The user can enter a 'seed' gene and choose one of the hand-curated microarray sets (**A**). yMGV computes the similarity between the expression profile of the seed and those of all other genes in this organism across the microarray set (**B**). Highly correlated (or anti-correlated) genes are selected (**C**) and a graphical representation shows their expression across microarrays of the set (**D**). Gene-specific links to external databases are provided, and the user can also post the whole gene list (**E**) to other databases in order to map them onto metabolic networks (KEGG) or the GO tree (SGD), or to try to find common cis-regulatory elements (RSA tools).

comparisons, a *S.cerevisiae* to *S.pombe* orthology table based on sequence similarity is stored in the database (the table was created by the Sanger Institute). The web interface allows users to retrieve genes based on their $\log_2$(ratio) thresholds in specified experiments. If the experiments are from different organisms, the corresponding orthology tables are used and only orthologs meeting specified thresholds are displayed (see Fig. 1 legend for details).

The evolutionary distance between *S.cerevisiae* and *S.pombe* (at least 400 million years), and the absence of a direct relationship between the sequence similarity and functional similarity of two proteins, influence the conclusions that can be drawn from cross-species comparison. Uncontrolled use of a module analysis based on orthology can yield misleading results. Accordingly, usage recommendations are associated with the module and can also be found in the Supplementary Material of this article. When used with discrimination, this tool should help the fission yeast community to easily take advantage of the huge amount of available information on the budding yeast transcriptome. yMGV is, to our knowledge, the first tool to allow this kind of comparison.

A tutorial explaining the use of the cross-species transcriptome comparison module is available at http://www.transcriptome.ens.fr/ymgv/tutorial/.

COMPENDIUM MODULE

Several years ago, it was shown that the application of various clustering algorithms to large microarray data sets can facilitate identification of biological meaningful groups (10,11). Since then, this approach has been used frequently and with great success, e.g. clustering was used to isolate interesting groups from a *S.cerevisiae* RNA data set of nearly 300 unrelated deletions or conditions (12). More recently, however, it has been shown that standard clustering methods are usually less effective when applied to large numbers of data sets that are biologically unrelated (13). Therefore, the microarray experiments in yMGV were hand curated and classified into 17 biologically coherent categories. We created a module that lists genes that are significantly co-expressed with respect to a user-selected reference gene according to a chosen metric and a chosen biological category (see Fig. 2). This proved to be very efficient for isolating genes co-regulated only in specific conditions.

A list of biological categories and some examples of usage are provided in the tutorial available at http://www.transcriptome.ens.fr/ymgv/tutorial/.

FUTURE DIRECTIONS

The major difficulty in maintaining a database like yMGV is data retrieval and curation. Thanks to the genomics community, standardization of microarray data sets (14) has facilitated the creation of central repositories for microarray data (15,16). We plan to incorporate deposited data sets into yMGV in order to maximize its utility to the biology community.

We also plan to add *cis*-regulatory elements to the yMGV output. This is essential, as phylogenetic footprinting has proved to be a very powerful technique that will become more and more efficient with increasing numbers of sequenced genomes, thus giving a more accurate description of the motifs involved in transcriptome regulation.

We are also planning to give users the option to upload their own data sets, and to use these data sets like any other data set in yMGV.

Finally, one of our long-term goals is to create a module that captures properties (expression regulation, GO annotations, *cis*-regulatory motif) from an input gene list and retrieves genes sharing similar or partially similar properties.

IMPLEMENTATION

The interface has been written in PHP and data are stored in a PostgreSQL relational database (logical scheme is available in Supplementary Material). yMGV uses data provided by external databases, namely GO descriptions from SGD (17) and GeneDB (www.genedb.org), and the orthology table provided by the Sanger Institute.

SUPPLEMENTARY MATERIAL

The Supplementary Material, available at NAR Online, contains: the database relational scheme, the yMGV data set contributors 2001–2003, the list of URLs to other databases and tools used in yMGV, and a description of limitations and potential problems associated with ortholog expression comparison.

ACKNOWLEDGEMENTS

The authors are grateful to the scientists who have supplied expression data and genome annotation (especially Valerie Wood), to Allegra Adele Petti for suggestions about the manuscript and to the following open source projects: Apache, Debian, PHP and PostgreSQL. The yMGV project was funded by the Programme Bioinformatique Inter-EPST-CNRS 2003. P.M. is supported by the French Therapeutical Research Association (AFRT) and the PhRMA foundation Center of Excellence in Integration of Genomics and Informatics (CEIGI).

REFERENCES

1. Gasch,A.P. (2002) Yeast genomic expression studies using DNA microarrays. *Methods Enzymol.*, **350**, 393–414.
2. Ulrich,R. and Friend,S.H. (2002) Toxicogenomics and drug discovery: will new technologies help us produce better drugs? *Nature Rev. Drug Discov.*, **1**, 84–88.
3. Wood,V. and Bahler,J. (2002) Website Review: How to get the best from fission yeast genome data. *Comp. Funct. Genomics*, **3**, 282–288.
4. Le Crom,S., Devaux,F., Jacq,C. and Marc,P. (2002) yMGV: helping biologists with yeast microarray data mining. *Nucleic Acids Res.*, **30**, 76–79.
5. Lyne,R., Burns,G., Mata,J., Penkett,C.J., Rustici,G., Chen,D., Langford,C., Vetrie,D. and Bahler,J. (2003) Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility and processing of array data. *BMC Genomics*, **4**, 27.
6. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
7. van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.
8. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
9. Marc,P., Devaux,F. and Jacq,C. (2001) yMGV: a database for visualization and data mining of published genome-wide yeast expression data. *Nucleic Acids Res.*, **29**, E63.
10. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.
11. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
12. Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
13. Gasch,A.P. and Eisen,M.B. (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.*, **3**, RESEARCH0059.
14. Spellman,P.T., Miller,M., Stewart,J., Troup,C., Sarkans,U., Chervitz,S., Bernhart,D., Sherlock,G., Ball,C., Lepage,M. *et al.* (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.*, **3**, RESEARCH0046.
15. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al.* (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
16. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
17. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al.* (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.